MOTION-BASED VIDEO OBJECT TRACKING IN THE COMPRESSED DOMAIN

Mark Ritch and Nishan Canagarajah

Department of Electrical & Electronic Engineering, University of Bristol, UK.

ABSTRACT

In this paper an algorithm for real-time unsupervised segmentation and tracking of a moving object is proposed. This is performed within the compressed domain using motion information only. Initial object segmentation is done using iterative rejection, taking advantage of its computational efficiency. The system seeks to overcome its disadvantages, namely a delay in object macroblocks appearing after consistency checking and non-identification of macroblocks containing object boundaries, by taking a model based approach to object tracking. The output of iterative rejection is used to update the model after tracking has taken place in each frame. Experimental results on a number of MPEG-2 encoded sequences demonstrate its effectiveness in identifying and tracking an object of interest from a compressed video stream and that the system is better than purely using iterative rejection as a segmentation method.

Index Terms— Compressed domain, Object tracking.

1. INTRODUCTION

Video object segmentation and tracking has many applications, including video indexing and retrieval, target recognition, surveillance applications, and object formation for content-based video standards such as MPEG-4 and MPEG-7. Although pixel domain segmentation can identify object boundaries with pixel accuracy, high computational complexity can result due to necessary decoding and the large number of pixels to be processed. Compressed domain techniques, on the other hand, only require partial decoding and can make use of the inherent encoded motion information. This is an advantage as most video is compressed into standardised formats for database storage and transmission.

Several approaches to video segmentation in the compressed domain have been proposed. Some methods utilise DCT information alone, such as Zeng et al. [1] where the DC sequence was extracted to compute the inter-frame difference and Sukmarg and Rao [2] used the DCT coefficients to perform region segmentation. Other methods segment using motion information and utilise the DCT information as a consistency measure. In [3] Gu used the thresholded difference in the motion vectors as a reliability check for interpolated macroblocks in B-frames. Babu et al. [4] used high DCT error energy and its variation to identify unreliable motion vectors. Both motion and DCT information was used by Porikli [5] in a volume-growing algorithm to create regions of similarity.

Although DCT information for each block may be approximated for P and B frames [6], this adds to the computational expense. Other methods utilise motion information only. Li et al. [7] proposed a spatiotemporal non-linear filter to detect objects and Favalli et al. [8] proposed a method for tracking objects using motion vectors, although the objects had to be initially specified manually. This tracking principle was used by Mezaris et al. [9] as a temporal consistency check for macroblocks labelled as 'object' by iterative rejection.

Iterative rejection is a computational efficient technique used to identify local motion. For compressed domain processing this algorithm is particularly attractive as motion vector information is readily available. However, a major drawback is that errors introduced in the encoder's motion estimation algorithm can lead to falsely identified local motion, particularly in homogenous regions. This problem was overcome to a certain extent in [9] by tracking macroblocks back through a number of frames. The greater the tracking depth, the more accurate this is, however, it leads to increasing delays before local motion is detected. In addition, less than optimal local motion identification occurs with increasing tracking depth and where a macroblock contains only a small amount of an object, typically seen at object boundaries. Conversely, greater than optimal local motion identification can arise where rapid or abrupt motion occurs and where object motion dominates the scene. Finally, the optimal threshold is also highly dependent on the content of the scene and segmentation performance can vary from frame to frame.

In this paper we propose a method to identify and track an object of interest within compressed MPEG-2 video using only motion information. The system is designed to detect interesting events taking place, such as a moving object appearing, and to track it without the need to decode the video. Iterative rejection is used as a basis for initial segmentation; however, the shortcomings identified above are addressed. This is done by adopting a novel model based approach, where the system maintains a binary model

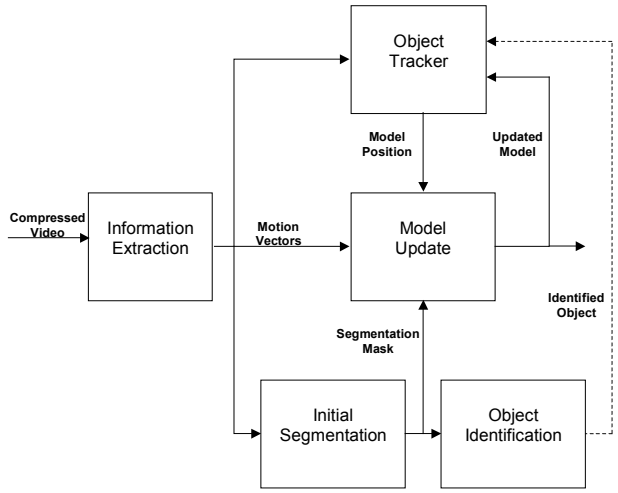

Figure 1 Block diagram of the proposed algorithm

of the object of interest in the form of a segmentation mask. A flow diagram of the method can be seen in Figure 1. In the next section the extraction of information required from the compressed video stream and the pre-processing applied are outlined. In Section 3 the process of iterative rejection is explained. The compressed domain tracker is presented in Section 4, and Section 5 details the process of updating the model. In Section 6 the system performance is evaluated and Section 7 provides some concluding remarks.

2. INFORMATION EXTRACTION FROM VIDEO

The system uses only motion information in the form of motion vectors which are extracted from the MPEG-2 compressed video stream. The motion vectors for P-frames can be used directly as they provide an indication of motion within the frame from the previous I or P-frame. Motion vectors for I-frames were estimated by averaging the motion vectors of adjacent P-frames [9], and those for Intracoded P-frame macroblocks were estimated by mean filtering the eight-point connected motion vectors, where available.

3. INITIAL SEGMENTATION

Initial segmentation is performed using iterative rejection, a method proposed for global motion estimation in [10]. The principle relies upon the proportion of the frame containing moving objects being significantly smaller than that containing background. The parameters of a global motion model are estimated and those macroblocks whose motion vector is larger than average are rejected. The process is repeated until no macroblocks are rejected. The result is that global, or camera, motion is estimated and those macroblocks affected by local, or object, motion are identified.

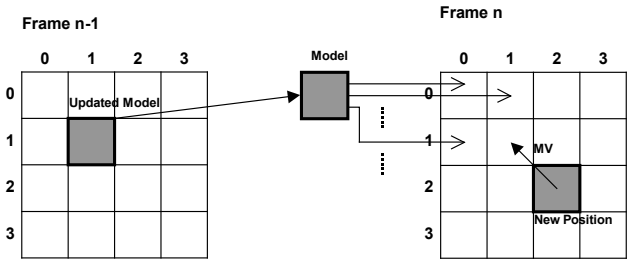

Figure 2 Block diagram illustrating compressed domain object tracking using motion vectors

4. OBJECT TRACKING USING MOTION VECTORS

The system derives and maintains a model for the object in the form of a segmentation mask, which is specified in terms of the associated positions of the macroblocks within the frame. To locate the most likely position of the current model within a new frame, the model is first abstracted from its position and overlaid in each macroblock position of the new frame. Figure 2 shows a block diagram of this process for a model of one macroblock.

The associated motion vector in each position is applied to give the position of the source of that macroblock in the previous frame. The distance between this position and that of the model in the previous frame is computed. The most likely position for the model in the current frame is deemed to be that position where the computed distance is minimum. For models of more than one macroblock the distance for each position in the frame is the average distance of all macroblocks in the model that overlap the frame.

5. UPDATING THE MODEL

Since the tracked object can change its shape from frame to frame, after tracking the model must be updated to allow it to expand and contract as necessary. The process begins by combining the segmentation mask of the tracked model with the iterative rejection mask and labelling regions using a recursive four-point connectivity algorithm [11]. The labelled region containing the tracked model is isolated and each macroblock within it falls into one of three categories.

Macroblocks that appear in both the tracked model and the iterative rejection mask are deemed to be consistent and automatically appear in the final updated model. Macroblocks that appear in the iterative rejection mask but not the tracked model become expansion candidates for inclusion in the model. Macroblocks that appear in the model but not the iterative rejection mask become contraction candidates for removal from the model.

If no model exists for a particular frame an object is extracted from the iterative rejection output after the temporal consistency check proposed in [8] is applied to each 'object' macroblock. This involves tracking the macroblock back through typically one or two previous

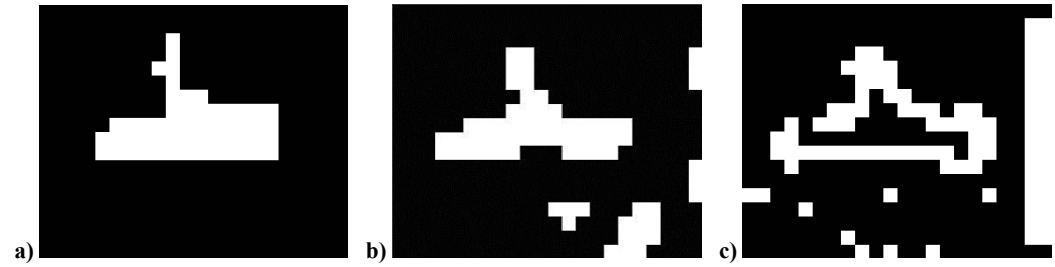

Figure 3 Frame 219 of 'coastguard' showing the binary mask of: a) the ground truth; b) the mean filtered motion vectors; and c) the spatial non-linear filtered motion vectors.

frames, and if the check fails the macroblock is removed from the iterative rejection mask.

5.1. Model Expansion

The model is allowed to expand when expansion candidate macroblocks have motion consistent with that of the model. This is achieved by estimating the parameters of a motion model using the updated model. If the difference between the local motion vector of the candidate macroblock and the motion vector predicted by the motion model is sufficiently small the model is allowed to expand. The distance between two motion vectors, *MV1* and *MV2*, is defined as the maximum of $|MV1_x - MV2_x|$ and $|MV1_y - MV2_y|$.

5.2. Model Contraction

The model is allowed to contract when contraction candidate macroblocks fail one of the two consistency checks outlined below:

5.2.1. Mean Filtered Motion Vectors

The local motion vector of the macroblock and those of the eight point connected neighbours are mean filtered. If the resulting magnitude of the macroblock in question is above a threshold the macroblock is deemed to be consistent and the model is not contracted. This is because the greater the strength of the surrounding local motion, the more likely the macroblock 'contains' at least part of an object and should not be removed. An example of the mean filtered motion vectors can be seen in Figure 3b.

5.2.2. Spatial Non-Linear Filtering

In [7] a spatial non-linear filter for use as a consistency check on motion vectors was proposed. The aim was to identify motion vectors that differed considerably from their neighbours, as they were less reliable at representing the underlying image motion. As a consistency check for model contraction we are positively attempting to identify this, the reason being that invalid contraction macroblocks are usually located at the edges of objects, particularly where the macroblock contains only part of the object of interest. The modified consistency measure is derived from the difference between the macroblock motion field, $\mathbf{d}(\mathbf{x},\mathbf{y})$, and the eight point connected neighbours, $\mathbf{N}(\mathbf{x},\mathbf{y})$, such that

$$C(x,y) = \sum_{(x',y')\in \mathbf{N}(\mathbf{x},\mathbf{y})} D\left(\mathbf{d}(\mathbf{x},\mathbf{y}), \mathbf{d}(\mathbf{x}',\mathbf{y}'), T_1\right),$$

where D is a binary function that gives 1 if the difference between $\mathbf{d}(\mathbf{x},\mathbf{y})$ and $\mathbf{d}(\mathbf{x}',\mathbf{y}')$ is greater than threshold T_1 and 0 otherwise. If the value of $C(x,y)$ is less than threshold T_2 then the macroblock is determined to be inconsistent with the edge of the model and the model is contracted. Typical threshold values are T_1=1.0 pixels and T_2=3. An example of the result of this filter can be seen in Figure 3c.

6. EXPERIMENTAL RESULTS

The method proposed was evaluated on a number of MPEG-2 compressed video sequences. Segmentation and tracking was performed on I and P-frames and a two-parameter translational motion model was used for both global motion estimation in iterative rejection and for model expansion.

Representative screen shots can be seen in Figure 4 for the coastguard, table tennis, hall and leopard sequences. From these it can be seen that the algorithm successfully identified the object of interest. In coastguard this is the smaller boat, which is tracked until it has left the scene, at which point the system identifies and tracks the larger boat. The same can be seen with both table tennis and hall, although with table tennis this occurs with the scene changes. Table 1 shows objective results for the coastguard sequence and Table 2 shows objective results for the leopard sequence, where the relevant thresholds were adjusted to maximise the F_1 measure. From this it can be seen that the spatial non-linear filter performs better as a method for model contraction, although some degradation in the precision occurs. In both cases it can be seen that the system performs better than purely segmenting the object of interest from the iterative rejection output with a temporal consistency of one previous frame.

7. CONCLUSION

In this paper an unsupervised model-based method for identifying and tracking an object of interest within compressed MPEG-2 video using only motion information

	Accuracy (%)			
	Recall	Precision	F_1 Measure	F_2 Measure
Mean filter	76.30	91.67	83.28	80.82
Spatial non-linear filter	85.21	82.94	84.06	84.44
Iterative rejection	80.36	63.13	70.71	73.66

Table 1 Performance Evaluation for 'Coastguard'

	Accuracy (%)			
	Recall	Precision	F_1 Measure	F_2 Measure
Mean filter	72.06	90.44	80.21	77.29
Spatial non-linear filter	76.80	88.57	82.27	80.36
Iterative rejection	70.73	86.05	77.64	75.19

Table 2 Performance Evaluation for 'Leopard'

has been presented. Iterative rejection is used to initially segment frames as it is computationally efficient and motion information can be sourced from the video stream without decompression. To overcome its disadvantages the system uses the iterative rejection output as a guide in identifying candidate macroblocks for inclusion into, and removal from, the model when it is updated. The decisions are made based on several measures of consistency using the motion vectors, which allow the model to expand and contract between frames. Experimental results on a number of sequences demonstrate its effectiveness in identifying and tracking an object of interest from a compressed video stream without the need to fully decode each frame, and that the system performed better than using iterative rejection alone as a segmentation method. Future work includes extending the system to identify and track multiple objects and to handle problems such as occlusion.

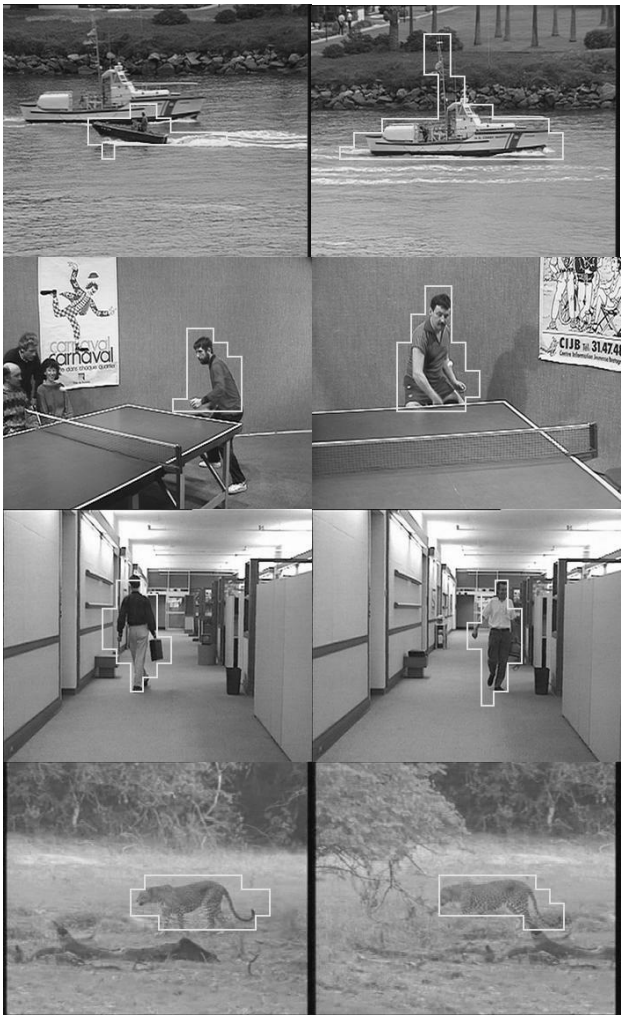

Figure 4 Final segmentation masks for the 'coastguard', 'table tennis', 'hall' and 'leopard' sequences.

8. REFERENCES

[1] W. Zeng, W. Gao, D. Zhao, "Automatic moving object extraction in MPEG video", *Proc. of the IEEE International Symposium on Circuits and Systems*, Vol. 2, pp. 524-7, 2003.

[2] O. Sukmarg and K.R. Rao, "Fast object detection and segmentation in MPEG compressed domain", *IEEE TENCON 2000*, Kuala Lumpur, Malaysia, Sept. 2000.

[3] L. Gu, "Scene analysis of video sequences in the MPEG domain", *Proc. of International Conference on Signal and Image Processing*, pp. 384-398, Las Vegas, October 1998.

[4] R.V. Babu, K.R. Ramakrishnan, and S.H. Srinivasan, "Video Object Segmentation: A Compressed Domain Approach", *IEEE Trans. CSVT*, Vol. 14, No. 4, pp. 462-474, April 2004.

[5] F.M. Porikli, "Real-time Video Object Segmentation for MPEG-encoded Video Sequences", *SPIE Conference on Real-Time Imagining VIII*, Vol. 5297, pp. 195-203, May 2004.

[6] B.L. Yeo and B. Liu, "On the extraction of DC sequence from MPEG video", *Proc. IEEE International Conference on Image Processing*, Vol. 2, pp. 260-263, 1995.

[7] J. Li, S. Kim, and J. Kuo, "Focus of attention (FOA) identification from compressed video for automatic target recognition (ATR)", *IEEE International Conference on Image Processing*, Washington, DC, pp. 508-511, Oct. 22-25, 1995.

[8] L. Favalli, A. Mecocci, and F. Moschetti, "Object tracking for retrieval applications in MPEG-2", *IEEE Trans. CSVT*, Vol. 10, No. 3, pp. 427-432, April 2000.

[9] V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, and M.G. Strintzis, "Real-Time Compressed-Domain Spatiotemporal Segmentation and Ontologies for Video Indexing and Retrieval", *IEEE Trans. CSVT*, Vol. 14, No. 5, pp. 606-621, May 2004.

[10] G.B. Rath, A. Makur, "Iterative Least Squares and Compression Based Estimations for a Four-Parameter Linear Global Motion Model and Global Motion Compensation", *IEEE Trans. CSVT*, Vol. 9, No. 7, pp. 1075-1099, October 1999.

[11] R. Jain, R. Kasturi, and B.G. Schunck, *Machine Vision*, McGraw Hill, 1995.